Computational Websites for Blended Statistics 371
Part 1: Chapters 1–9 of the Course Notes

Robert L. Wardrop, Emeritus Professor
Department of Statistics, UW–Madison
August 24, 2013

The Course Notes contain many references and links to computational websites that are needed in this course. Students have told me that it would be convenient to have all of these sites available in one place and this document is part of my response to this request. Part II of this document, which covers Chapters 10–22 of the Course Notes, will be available soon.

These documents will be posted on Learn@UW and also on my website:

http://www.stat.wisc.edu/~wardrop/

An advantage of the latter posting is that from it you can use the links below to access the sites directly.

This document is much longer than it needs to be; the teacher in me could not resist providing explanations and examples of the uses of these sites, even though similar explanations and examples are provided in the Course Notes.

1 Chapter 1

The only site we access in Chapter 1 is:

http://vassarstats.net.

This is an extremely versatile site and we will use it often in this course.

This vassarstats site almost always provides more output than we desire. In Chapter 1, we use it simply to obtain the mean, variance and standard deviation of a set of data.

The left margin of the website presents a list of links; click on t-Tests & Procedures. This takes you to a new page; click on Single Sample t-Test. The new page contains a box for data entry; I will illustrate its use with the chicken data from Dawn’s study of Bob: 1, 3, 4, 5, 5, 6, 6, 6, 7 and 8. I type these numbers into the date entry box; one number per line. I click on calculate and obtain the following statistics:

- $n = 10$.
- variance (inferential): $s^2 = 4.1$.
- standard deviation (inferential): $s = 2.0248$.
- sample mean: $\bar{x} = 5.1$. 

The other output provided by vassarstats can be ignored safely.

There is a website that will draw a dotplot of a set of data. You won’t need to use it in this course, but I include it for completeness. It is:

http://rossmanchance.com/applets/DotPlotApplet/DotPlotApplet.html

If you are interested in this website, I suggest you try it with the chicken data from Dawn’s study.

2 Chapter 2

No new websites are needed in Chapter 2. For completeness, here is a website that will draw a kernel density for a data set:

http://www.wessa.net/rwasp_density.wasp#output

You will not need to use this site in this course. But if you want to check it out, proceed as follows.

Near the top of the page you will see a box labeled Data with a long listing of data below it: 112, 118, 132, . . . . To see what this site produces, keep this default data set and click on Compute at the bottom of the box. After waiting a few seconds, scroll down and see the different densities that have been computed. (In the Course Notes I use a version of the Gaussian Kernel.)

3 Chapter 3

One website is introduced in Chapter 3:

http://www.randomizer.org/form.htm

In Chapter 3 we use this site to obtain assignments using the process of randomization. I will illustrate the use of this site for Dawn’s study of her cat Bob. Recall that Dawn decided to perform a balanced CRD with a total of $n = 20$ trials.

When you call this site, you will note that there are seven boxes awaiting your input:

1. How many sets of numbers do you want to generate? If you are a researcher seeking an assignment to use for your CRD, use the default value 1. If for some reason—for example, a homework assignment—you want to generate $k > 1$ assignments, then enter the integer value of $k$ in the box.

2. How many numbers per set? Enter the value of $n_1$ in the box, where, recall, $n_1$ is the number of units (trials, subjects) to be assigned to the first treatment. For Dawn’s study, $n_1 = 10$.

3. Number range (e.g., 1-50): We always label our units 1, 2, 3, . . . $n$; thus, leave the default value ‘1’ in the From box.

4. Number range (e.g., 1-50): Enter the value $n$ in the To box. For Dawn’s study, $n = 20$.

5. Do you wish each number in a set to remain unique? Leave the default answer, Yes, in the box.

6. Do you wish to sort the numbers that are generated? Reasonable people can disagree with me, but following the method in the Course Notes, we need to sort the values. Thus, change from the default No to the option Yes: Least to Greatest.
7. How do you wish to view your random numbers? Leave the default value, Place Markers Off in the box.

Once you are satisfied with the entries in the seven boxes, click on the box Randomize Now!. I did so and obtained the following output:

\[2, 5, 7, 8, 9, 13, 14, 17, 19 \text{ and } 20.\]

This assignment places the listed trials on the first treatment and the remaining trials (1, 3, 4, 6, 10, 11, 12, 15, 16 and 18) on the second treatment. Alternatively, it sometimes is useful to view this information as follows:

<table>
<thead>
<tr>
<th>Trial</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treat.</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

4 Chapter 4

No new websites are needed in Chapter 4.

5 Chapter 5

In Chapter 5, we revisit the vassarstats website:

http://vassarstats.net

We use this site to obtain approximate P-values for the test of Chapter 5—the comparison of means. After accessing this site, select Miscellanea from the list of links in the left margin; it is the penultimate entry. On the new page, click on Resampling Probability Estimates ... and follow the instructions, which I will illustrate with Dawn’s study of Bob.

1. Enter the value of \(n_1\) in the box for the size of sample A; \(n_1 = 10\) for Dawn’s study. Click on OK.

2. Enter the value of \(n_2\) in the box for the size of sample B; \(n_2 = 10\) for Dawn’s study. Click on OK.

3. On the new page, find the Data Entry section, enter the data and click on Calculate.

4. Much of the output may be ignored, but check to make sure the means are correct. They are 5.1 and 2.9 for Dawn’s data, which match my earlier results; thus, you have a partial check that the data were entered correctly!

5. Move down to the box Resample x 1000. To obtain a computer simulation with \(m = 10,000\) runs, click on this box exactly 10 times. I did this for Dawn’s data and obtained the following output next to the label \(P_{\text{resampling}}:\)

- one-tailed \(0.0181\);
- two-tailed \(0.0368\).

Below I explain how to interpret these two proportions.

Let’s begin with the easy output. The approximate (the site calls it estimated) P-value for the alternative \(\neq\) appears next to two-tailed. For my simulation of Dawn’s data, the entry is 0.0368; if you repeat my work you will likely obtain a different answer. (For example, I performed a second 10,000 run simulation study and obtained 0.0370.)
Ideally, *vassarstats* would provide approximate P-values for each of the remaining alternatives, > and <. Sadly, it does not; it gives only one of these P-values. For my simulation results it gave 0.0181 which is for the alternative >. How do I know this? Well, the rule is pretty simple to remember:

- If the observed value of the test statistic $u$ is a positive number, then the *vassarstats* approximate P-value is for the alternative $>$. 
- If the observed value of the test statistic $u$ is a negative number, then the *vassarstats* approximate P-value is for the alternative $<$. 
- If the observed value of the test statistic $u$ is zero, then the *vassarstats* approximation should not be used. All we know is that the exact P-value for both $>$ and $<$ exceeds 0.5000.

If you want an approximate P-value for the alternative not covered by *vassarstats* you may proceed as follows, as illustrated with my results reported above.

Recall that the observed value of the test statistic for Dawn’s data is $u = 2.2$. Thus, the formulas for the one-sided P-values are:

- For $>$: The P-value equals $P(U \geq 2.2)$.
- For $<$: The P-value equals $P(U \leq 2.2)$.

In addition, we know that the sum of the exact P-values for $<$ and $>$ for any study is

$$P(U \leq u) + P(U \geq u) = 1 + P(U = u) > 1.$$ 

Thus, given that I approximate $P(U \geq 2.2)$ for Dawn’s data by 0.0181, it makes sense to say that

$$P(U \leq 2.2)$$ is at least (approximately) $1 - 0.0181 = 0.9819$.

In words, it is a really huge P-value!

### 6 Chapter 6

Chapter 6 presents the sum of ranks test for comparing two treatments and a numerical response. If the number of possible assignments is small, then finding the exact sampling distribution might be plausible. Otherwise, you can use a computer simulation experiment to obtain an approximate P-value for the test.

Here is some good news. You can use the website we used last chapter:

[http://vassarstats.net](http://vassarstats.net).

Instead of entering the data, you must enter the ranks and then follow the steps exactly as outlined above for Chapter 5. I will illustrate the method for Dawn’s study.

As shown in Chapter 6 of the Course Notes, the ranks for Dawn’s data are given below.

<table>
<thead>
<tr>
<th>Treat. 1:</th>
<th>3</th>
<th>7.5</th>
<th>10.5</th>
<th>13</th>
<th>13</th>
<th>16</th>
<th>16</th>
<th>16</th>
<th>18.5</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treat. 2:</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>7.5</td>
<td>7.5</td>
<td>7.5</td>
<td>7.5</td>
<td>10.5</td>
<td>13</td>
</tr>
</tbody>
</table>

I entered these ranks into *vassarstats*. First, the site gave me the mean ranks: 13.35 for treatment 1 and 7.65 for treatment 2. Next, I performed a simulation study with $m = 10,000$ runs. My results were: the one-tailed approximate P-value is 0.0150 and the two-tailed approximate P-value is 0.0305. Because the mean rank on treatment 1 is larger than the mean rank on treatment 2, the one one-tailed P-value is for the alternative $>$. 

4
Chapter 7 introduces the family of Normal curves. As you will see in later chapters of the Course Notes, Normal curves play an important role in many statistical analyses. Recall that a Normal curve is determined by the values of two parameters: its mean, denoted by $\mu$, which can be any real number; and its standard deviation, denoted by $\sigma$, which can be any positive real number. The family of Normal curves is, indeed, a very large family!

The terminology

\[ N(\mu, \sigma) \]

means the Normal curve with mean $\mu$ and standard deviation $\sigma$. The most important member of the family of Normal curves is its matriarch, the $N(0,1)$ curve, also called the standard Normal curve.

In the old days (walking to school, through snow drifts, uphill in both directions) when we had a question about the area under a Normal curve, we first would standardize the question so that it could be answered by referring to a table of areas under the standard Normal curve. Fortunately for you, we are now in the internet age; as a result, no standardization is necessary. You can find your answer by using the following website:

http://davidmlane.com/hyperstat/z_table.html

Please access this site to follow my explanation below.

The site begins by presenting a graph of the standard Normal curve. Immediately below the curve—easy to overlook—is a pair of options, called: Area from a value; and Value from an area. The first of these is the default option and I will begin by explaining it.

The site asks you to specify the Normal curve that interests you, with the standard Normal curve as the default. For the sake of illustration, let’s use the $N(100,15)$ curve. I type 100 in the box next to Mean and 15 in the box next to SD. (The shading under the curve changes when I do this, but don’t worry about it.) The site will answer four types of questions, corresponding to the options:

Above; Below; Between; and Outside.

Our main interest rests with Above and Below, but I might refer to the other options on occasion. Let’s start with Above.

First, note that Above is synonymous with greater than or to the right of. Suppose that I want the area under the $N(100,15)$ curve to the right of 119.5. I select the option Above (the default) and enter 119.5 in the box to the right of it. I click on the box Recalculate and the answer, 0.0912, appears in the box next to Area (probability). You might note that clicking on Recalculate changes the graph and its shaded area. Not being a particularly visual person, I usually ignore this.

Next, suppose that I want the area under the $N(100,15)$ curve to the left of 86.5. I select the option Below, enter 86.5 in its box and click on Recalculate; the answer is 0.1841.

The option Value from an area will be covered in Chapter 9 below.

Now that you know about Normal curves, you can use math theory to obtain an approximate P-value for the sum of ranks test. As discussed in the Course Notes, first you calculate the values of: $r_1$; the mean of $R_1$; and the standard deviation of $R_1$. Given these values, the approximate P-value can be obtained using

http://davidmlane.com/hyperstat/z_table.html

as described above. Alternatively, you can save yourself from hand calculations by using vassarstats.

Go to the website:
Click on Ordinal Data in the left margin. Next, click on the Mann–Whitney Test. Now things get tricky. For an unbalanced study, the site requires \( n_1 \) to be larger than \( n_2 \). To accommodate this restriction, you may need to renumber your treatments and, if you do, remember that this will reverse a one-sided alternative—for example, \( > \) will become \( < \). The alternative \( \neq \) is not affected by renumbering the treatments.

I will illustrate the use of the site with Reggie’s study of darts, which was introduced in the Chapter 1 Homework. Reggie’s study was balanced, with \( n_1 = n_2 = 15 \); thus, there is no need to renumber the treatments—treatment 1 will remain the shorter distance.

Back to the site. I enter the two sample sizes and am taken to a page which offers me three options for entering my data. First, I can Import (cut-and-paste) my data, which would be the most painless method. Second, I can enter my observations one-at-a-time. Finally, I can enter my ranks, one-at-a-time. I will use the second method.

I entered the data and clicked on Calculate from Raw Data. The site yields quite a lot of output, most of which we will ignore.

The site gives the mean ranks of the data: 19.3 for treatment 1 and 11.7 for treatment 2. Thus, based on ranks, Reggie’s observations on treatment 1 are larger, as a group, than his observations on treatment 2. As before, this tells us that the one-sided P-value that vassarstats provides—denoted by \( P_1 \) = 0.0091—is for the alternative \( > \). Finally, the site gives \( P_2 = 0.0183 \), which is the approximate P-value for the two-sided alternative, \( \neq \). Of course, because the approximating Normal curve is symmetric, the latter P-value equals twice the former, except for round-off error.

Recall, as discussed in the Course Notes, this vassarstats site uses the uncorrected (for ties) formula for the standard deviation of \( R_1 \). As a result, the P-values it reports are a bit larger than what one would obtain by using the accurate formula. Using Minitab, I found that corrected P-values are 0.0179 for \( \neq \) and 0.0089 for \( > \); in words, whether or not one corrects for ties makes virtually no difference in the approximate P-values.

In the remainder of this section, I will show how to obtain both the vassarstats and Minitab results by hand. If this topic does not interest you, feel free to jump ahead to the section on Chapter 8.

For Reggie’s study, \( n_1 = n_2 = 15 \). Thus, the mean of \( R_1 \) is:

\[
\mu = n_1(n + 1)/2 = 15(31)/2 = 232.5.
\]

The uncorrected (no-ties) variance of of \( R_1 \) is:

\[
\sigma^2 = n_1n_2(n + 1)/12 = 15(15)(31)/12 = 581.25, \text{ giving } \sigma = 24.109.
\]

In the homework for Chapter 6, we found that the observed value of \( R_1 \) for Reggie’s data is \( r_1 = 290 \).

Thus, the P-value is \( P(R_1 \geq 290) \), which we convert, via the continuity correction, to \( P(R_1 \geq 289.5) \). The area under the N(232.5,24.109) curve to the right of 289.5 is 0.0090, which agrees with the vassarstats value of 0.0091, except for round-off error.

Furthermore, for Reggie’s data, there are six instances of \( t_i = 2 \) and one instance of \( t_i = 4 \). All other values of \( t_i \) are 1. (Trust me on this.) Thus,

\[
\sum (t_i^3 - t_i) = 6(8 - 2) + 1(64 - 4) = 36 + 60 = 96.
\]

Thus, the corrected variance is:

\[
\sigma^2 = 581.25 - \frac{15(15)(96)}{12(30)(29)} = 581.25 - 2.07 = 579.18,
\]

6
giving $\sigma = 24.067$. The area under the $N(232.5, 24.067)$ curve to the right of 289.5 is 0.0089, which agrees with the Minitab answer to four digits of precision.

8 Chapter 8

When the response for a CRD is a dichotomy, then the test that compares proportions is called Fisher’s test. The exact P-value can be obtained by using the site:

http://www.langsrud.com/fisher.htm

9 Chapter 9

Our Normal curve area website:

http://davidmlane.com/hyperstat/z_table.html

can be used to find the approximate critical region for the sum of ranks test. This technique is a bit complicated; thus, I will give a detailed example of its use.

Suppose that we have a balanced CRD with a numerical response and a total of $n = 100$ trials. According to the formulas given in Chapter 7 of the Course Notes, the mean of $R_1$ is

$$\mu = n_1(n + 1)/2 = 50(100)/2 = 2500,$$

and, assuming there are no ties in the 100 observed values, the standard deviation of $R_1$ is

$$\sigma = \sqrt{n_1 n_2 (n + 1)/12} = \sqrt{50(50)(101)/12} = \sqrt{21041.67} = 145.06.$$

Thus, I will use the $N(2500, 145.06)$ curve to approximate the sampling distribution of the test statistic $R_1$.

Now suppose that I have chosen the alternative $>$ and I want the critical region for the test with $\alpha$ equal to (approximately) 0.05. Thus, I want to find the number $c$ with the property that

$$P(R_1 \geq c) = 0.05).$$

We proceed as follows.

Go to:

http://davidmlane.com/hyperstat/z_table.html

Choose the option Value from an area and enter the target probability (0.05) in Area as well as the values for Mean (2500) and SD (145.06). The site yields the answer 2738.654 next to Above. Thus, our value for $c$ using the Normal curve approximation is $c = 2738.654$ and the critical region is $R_1 \geq 2738.654$.

The above answer is fine, but I will refine it a bit. Because there are no ties in the data, there will be no fractional ranks and thus, the observed value of the test statistic must be an integer. Thus, having a fractional component to the critical value, while not literally incorrect, is inelegant. Thus, let me round my value of $c$ to 2739. Thus, my critical regions is $R_1 \geq 2739$, which looks neater. Again, I could stop here, but I want to return to the Normal curve area website once more.

I want to calculate the value of $\alpha$ for my critical region; i.e., I want $P(R_1 \geq 2739)$. I use the continuity correction to change this expression to $P(R_1 \geq 2738.5)$ and return to the Normal curve area website to obtain an approximation to this probability, as outlined below.
Select the option *Area from a value*;

Enter 2500 for *Mean* and 145.06 for *SD*;

Keep the default option *Above* and enter 2738.5 in its box;

Finally, click on *Recalculate* and read the answer: 0.0501 is the approximate value of $\alpha$ for the critical region $R_1 \geq 2739.$